Reconhecimento online de gestos dinâmicos para ambientes interacionais multicâmeras

Clebeson Canuto [*] **Luiz Carlos Cosmi** [*] **Alexandre Pereira** [**]
Jorge Samatelo [*] **José Santos-Victor** [***] **Raquel Vassallo** [*]

[*] *Universidade Federal do Espírito Santo, Vitória - ES, Brasil, (e-mail: {clebeson.canuto, luizcarloscosmifilho}@gmail.com, jorge.samatelo@ufes.br, raquel@ele.ufes.br).*
[**] *Instituto Federal do Espírito Santo Guarapari-ES, Brasil, (e-mail: alexandre.carmo@ifes.edu.br)*
[***] *Intituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, (e-mail: jasv@isr.tecnico.ulisboa.pt)*

Abstract: This work proposes an online dynamic gesture recognizer that can be used in an interactive multicamera environment. The proposal consists of reconstructing the three-dimensional skeleton of a user present in the scene, using a temporal segmentation model to segment the skeleton sequences that contain gestures and classifying them into one of the possible classes. Thus, the main contributions in this work are three-fold: (*i*) a model able to segmenting a temporal flow of 3D skeletons as being gestures or non-gestures; (*ii*) a model that allows classifying a sequence of 3D skeletons movements as belonging to one of the possible gesture classes; and (*iii*) a system that unites the segmentation and classification models, in order to allow the recognition of online gestures in a multicamera interactive environment. To evaluate the proposal, two gesture datasets were acquired in two distinct interactional spaces. After the experiments, the proposed spotting model obtained an average accuracy over the test data set of 82%, and the classifier model obtained 72.40%. The two models together reached a Jaccard index of 0.76. Considering the processing time, each observation required an average time of 70ms when executed in GPU. Thus, the proposed solution is considered effective in online dynamic gesture recognition for multicamera environments.

Resumo: Este trabalho propõe um reconhecedor de gestos dinâmicos que pode ser utilizado em um ambiente interacional multicâmera de modo *online*. A proposta consiste em reconstruir o esqueleto tridimensional de um usuário presente na cena, utilizar um modelo se segmentação temporal para segmentar as sequências de esqueletos que contenham gestos e classificá-las em uma das possíveis classes. Dessa forma, este trabalho traz três contribuições principais: (*i*) um modelo capaz de segmentar temporalmente um fluxo de esqueletos 3D como sendo gestos ou não-gestos; (*ii*) um modelo que permite classificar uma sequência de movimentos de esqueletos 3D como pertencentes a uma das classes de gestos possíveis; e (*iii*) um sistema que une os modelos de segmentação e classificação, de maneira a possibilitar o reconhecimento *online* de gestos em um ambiente interacional multicâmera. Para valiar a proposta foram utilizada duas bases de gestos capturadas em dois espaços interacionais distintos. Após os experimentos, o modelo de segmentador proposto obteve uma acurácia média sobre o conjunto de dados de teste de 82% e o modelo classificador obteve $72,40\%$. Os dois modelos em conjunto alcançaram um índice de Jaccard de 0.76. Considerando o tempo de processamento, cada observação requereu um tempo médio de 70ms quando executado em GPU. Dessa forma, julga-se a solução proposta efetiva no reconhecimento *online* de gestos dinâmicos para ambientes interacionais multicâmeras.

Keywords: Dynamic Gesture Recognition; Interactive Environment, Intelligent Space; Multicamera Environment; Deep Learning.

Palavras-chaves: Reconhecimento Gestos Dinâmicos; Ambiente Interacional, Espaço Inteligente; Ambiente Multicâmera; Aprendizagem Profunda.

1. INTRODUÇÃO

Um espaço inteligente pode ser entendido como um ambiente capaz de tomar decisões por meio de dados coletados de vários sensores e dispositivos interconectados (Lee et al., 1999). Em um espaço inteligente, são diversos os possíveis agentes coletores de dados, por exemplo, robôs, computadores, sensores de presença, câmeras, dentre outros. Alguns destes agentes, como é o caso dos robôs, também são atuadores, recebem instruções do espaço sobre as tarefas a serem executadas. Dentro do ambiente monitorado, um usuário pode interagir diretamente com os dispositivos, bem como com o próprio ambiente. Dessa forma, como as decisões são tomadas pelo espaço, ele deve ser capaz de prover a infraestrutura necessária para a que interação ocorra efetivamente, independente de para quem ela foi dirigida. Dessa maneira, para facilitar a utilização das aplicações em diferentes locais, um ambiente interacional deve satisfazer dois requisitos principais: utilizar uma interface de interação intuitiva e utilizar sensores comuns aos mais diversos ambientes.

Dentre as interfaces de interação, os gestos dinâmicos são considerados uma das mais intuitivas (Santos et al., 2016). Um gesto dinâmico pode ser definido como um conjunto de pequenos movimentos, comumente realizados pelos braços, mas não restritos a eles, que visam a comunicação entre indivíduos. Um gesto pode ser compreendido como sendo a evolução da pose (configuração das juntas) de um indivíduo em um intervalo de tempo. Sendo assim, um sistema automático que objetive o reconhecimento de gestos deve considerar essa definição e utilizar aqueles sensores que melhor capture esses tipos de características.

Desde o advento do *Microsoft Kinect* 360, em meados de 2011, o reconhecimento de gestos dinâmicos pode ser executado de maneira mais efetiva. Isso é devido ao fato de que o sensor de profundidade presente no kinect possibilita a extração dos esqueletos de pessoas presentes em seu campo visual. Dessa forma, é possível utilizar modelos relativamente simples para reconhecer gestos dinâmicos, uma vez que o esqueleto fornece uma representação compacta do movimento corporal de uma pessoa. O problema com o uso desse tipo de sensor é que mesmo sendo de fácil aquisição e com um custo acessível, a quase totalidade dos ambientes interacionais onde um reconhecedor de gestos pode ser utilizado possuem câmeras RGB como principais sensores. Com isso, modelos dependentes do kinect possuem uma restrição natural de uso, uma vez que tais ambientes com câmeras convencionais têm apenas as imagens RGB como a principal fonte de informação para o reconhecimento de gestos.

Adaptando a definição anterior para a área de visão computacional (análise automática de imagens), um gesto dinâmico pode ser entendido como a evolução da pose (configuração das juntas) de uma pessoa em uma sequência de imagens. Dessa maneira, um modelo que pretende resolver esse problema enfrenta diferentes desafios, tanto na extração de características espaciais (obtenção da pose e/ou movimento em cada imagem), quanto na análise da evolução de cada característica (relacionar as características temporalmente). Isso levanta uma barreira nas pesquisas de reconhecimento de gestos em vídeos, e direciona esforços para modelos baseados em esqueletos.

O problema com o reconhecimento de gestos dinâmicos é ainda maior dentro do escopo dos espaços inteligentes multicâmeras. Mesmo que as várias câmeras possam ajudar na diminuição de características oclusas, elas aumentam significativamente a necessidade de processamento do classificador e ainda geram um problema referente à decisão de que câmera deve ser utilizada no reconhecimento.

Em uma situação de reconhecimento *online* esses problemas são ainda mais presentes, uma vez que não se tem a informação do tamanho de cada sequência que representa um gesto. Além disso, um mesmo gesto pode ser realizado com mais ou menos imagens a depender do usuário. Ou seja, por um lado mais câmeras aumenta a área monitorada pelo espaço ao mesmo tempo que reduz os problemas com oclusão, por outro lado, quanto mais câmeras, mais complexo pode se tornar o processo de reconhecimento de gestos. Sendo assim, objetivando o uso de ambientes multicâmeras para a interação natural, Queiroz et al. (2018) propôs um método que usa as imagens de N câmeras de um espaço inteligente calibrado para reconstruir o esqueleto tridimensional de cada indivíduo presente no mesmo. Com essa solução, é possível utilizar o potencial dos reconhecedores de gestos baseados no Kinect mediante múltiplas câmeras.

Mesmo com a possibilidade de utilização de esqueletos 3D, alguns trabalhos utilizam modelos baseados em sequências fixas de imagens (Escobedo-Cardenas and Camara-Chavez, 2015; Liu et al., 2019), o que acaba gerando uma dependência de como os gestos devem ser executados, e por conseguinte limita a sua naturalidade de execução. Um problema similar ocorre com as abordagens baseadas em regras (Santos et al., 2015), que apresentam uma forte dependência da percepção do projetista. Uma ligeira mudança na realização de um gesto pode prejudicar o seu reconhecimento. Sendo assim, uma abordagem baseada em um modelo de aprendizado automático, orientado à tarefa de classificação de gestos e que não limitasse o tamanho da sequência de entrada, superaria as restrições dos modelos de sequências fixas e os baseados em regras.

Dessa maneira, este trabalho propõe um reconhecedor de gestos *online* que possa ser utilizado em um espaço interacional multicâmera. A proposta consiste em utilizar o método apresentado em Queiroz et al. (2018) para extrair o esqueleto 3D de um usuário presente na cena, fornecer o esqueleto para um modelo que o classificará como sendo gesto ou não-gesto. Isto possibilitará a segmentação temporal de sequências contendo gestos, as quais deverão alimentar um outro modelo responsável por classificá-las em um dos gestos possíveis.

⋆ Os autores agradecem à CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pelo suporte financeiro dado através da Bolsa de Doutorado concedida ao primeiro autor. À FAPES - Fundacão de Amparo à Pesquisa e Inovação do Espírito Santo, pelo suporte financeiro dado através da bolsa de apoio técnico concedida ao segundo autor (Projeto 577/2018). Também à NVIDIA pela doação da placa gráfica Titan V utilizada nos experimentos.

Em resumo, as principais contribuições deste trabalho são as seguintes:

- um modelo capaz de segmentar temporalmente um fluxo de esqueletos 3D como sendo gestos ou não-gestos;
- um modelo que permite classificar uma sequência de movimentos de esqueletos 3D como pertencentes a uma das classes de gestos possíveis; e
- um sistema que une os modelos de segmentação e classificação citados acima, de maneira a possibilitar o reconhecimento *online* de gestos em um ambiente interacional multicâmera.

Dessa maneira, para um melhor entendimento, as próximas seções abordarão, respectivamente: os trabalhos relacionados (Seção 2), a descrição completa da proposta (Seção 3), como os experimentos foram preparados e executados (Seção 4), quais os resultados obtidos (Seção 5) e quais as conclusões e os trabalhos futuros (Seção 6).

2. TRABALHOS RELACIONADOS

Com o aumento no desenvolvimento da robótica e dos sistemas computacionais, a necessidade de uma interação intuitiva é uma realidade inevitável. Nesse sentido, o reconhecimento de gestos dinâmicos é uma área de pesquisa que vem ganhando cada vez mais importância, porém ainda carece de muito desenvolvimento.

Dentre os trabalhos que abordam esse assunto, a maioria foca no reconhecimento de gestos *offline*, ou seja, reconhecer um gesto presente em uma sequência de observações previamente segmentadas. Em Santos et al. (2020) é apresentado um reconhecedor de gestos baseado em imagens RGB. Nele, os autores propuseram representar o movimento presente em uma sequência de *frames* em apenas uma imagem, chamada de *star RGB*. Essa imagem resultante é dada como entrada para um classificador baseado em duas Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Networks*), que são fundidas por um mecanismos baseado em atenção (*Soft-attention*). O resultado da fusão alimenta uma MLP (*Multilayer Perceptron*), responsável pela classificação do gesto presente na sequência de imagens de entrada. O trabalho trás ótimos resultados para três bases de dados distintas, porém, assim como discutido pelos autores, para aplicações em tempo real é necessária uma etapa de segmentação prévia, uma vez que o modelo precisa que cada sequência de entrada possua apenas um gesto.

Em Cao et al. (2015), os autores utilizam um HMM (*Hidden Markov Model*) a fim de reconhecer gestos em uma sequência de esqueletos 3D. Já Wang and Wang (2017) dividiram o esqueleto 3D em 5 partes principais (pernas, braços e tronco) e forneceram a sequência correspondente a cada uma dessas partes para uma rede neural recorrente do tipo LSTM (*Long-Short Term Memory*). As saídas das 5 LSTM eram então concatenadas e dadas como entrada para uma outra LSTM que classificava a sequência de esqueletos como pertencentes a uma das classes de gestos.

Chai et al. (2016) propuseram uma abordagem capaz de reconhecer gestos dinâmicos de modo *online*. Primeiro, realiza-se uma segmentação temporal no vídeo contendo os gestos. Para isso, são extraídas características das mãos das pessoas por meio de uma Fast-RCNN (Girshick, 2015), arquitetura utilizada no reconhecimento de objetos, e do descritor HOG (*Histogram of Oriented Gradients*). Essas características alimentam uma rede neural recorrente, responsável por classificar cada *frame* de entrada como sendo gesto ou não-gesto. Dessa maneira, as sequências marcadas como sendo gesto alimentam um classificador baseado numa rede LSTM. Nessa aboradagem, os autores utilizam tanto imagens RGB quanto imagens de profundidade.

Em Neverova et al. (2015) é apresentado uma abordagem similar ao proposto em Chai et al. (2016). Aqui, os autores extraem características das configurações das juntas de um usuário (velocidades, ângulos e acelerações) e as fornecem para um segmentador temporal baseado numa MLP com apenas uma camada escondida. Similarmente ao trabalho anterior, as sequências segmentadas são então passadas para uma arquitetura que utiliza, além do esqueleto, imagens RGB, de profundidade e áudio para classificar o gesto presente na sequência. Apesar das duas abordagens oferecerem soluções para o reconhecimento de gestos em tempo real, as mesmas utilizam dados multimodais e necessitam de grandes conjuntos de dados para serem treinadas. Sendo assim, caso fossem utilizadas em um ambiente multicâmera, a complexidade de seus modelos cresceria consideravelmente, uma vez que, mesmo sendo multimodais, as informações utilizadas são adquiridas de um único sensor, mais especificamente, um Kinect 360.

Considerando agora os ambientes interacionais, trabalhos que abordam o reconhecimento *online* de gestos para esse tipo de ambiente são escassos. Em Santos et al. (2017), os autores modelam um comportamento interacional capaz de interagir com um robô móvel por meio de gestos dinâmicos. Apesar dos gestos serem reconhecidos em tempo real, o reconhecimento é baseado em um conjunto de regras pré-definidas que se baseiam na localização de determinadas juntas do usuário. Esse tipo de reconhecedor é altamente propício a falsos positivos quando da existência de pequenas variações na maneira como os gestos são executados por diferentes usuários. Além disso, depende exclusivamente da experiência do projetista; não é aprendido por meio da análise de dados. Em conclusão, reconhecedores baseados em regras limitam a sua utilização a ambientes interacionais específicos e cuidadosamente projetados. Outras abordagens semelhantes também podem ser vistas em Zhao et al. (2013); Santos et al. (2015); Saleme et al. (2017).

Mesmo com os problemas apresentados nos trabalhos anteriores, vê-se a importância do uso de esqueletos para o reconhecimento de gestos, ao mesmo tempo que o uso de sensores de profundidade pode limitar a escalabilidade da solução. Sendo assim, visando a interação em espaços inteligentes multicâmeras, Queiroz et al. (2018) propuseram um método capaz de reconstruir tridimensionalmente as juntas dos esqueletos detectados em imagens capturadas por um sistema multicâmeras calibrado. O método proposto pelos autores pode ser explicado em cinco principais etapas: primeiro, utiliza-se o *openpose* (Cao et al., 2018) para detectar as juntas dos usuários presentes em cada imagem; segundo, os esqueletos formados pelas juntas recebem um identificador único e são então associados às respectivas câmeras nas quais foram identificados; em seguida, utiliza-se de geometria epipolar para encontrar a

correspondência entre os esqueletos de câmeras diferentes; dessa forma, por meio de uma busca em grafo, agrupam-se as correspondências a fim de evitar redundâncias; finalmente, a reconstrução tridimensional das juntas é feita utilizando suas coordenadas e os parâmetros de calibração das câmeras.

Após essa análise, percebe-se que são poucos os trabalhos que oferecem soluções para o reconhecimento de gestos dinâmicos em ambientes interacionais, e os que assim o fazem, possuem um escopo limitado, o que impossibilita a sua utilização em ambientes mais diversificados. Uma outra percepção é que diferentes trabalhos utilizam o esqueleto como fonte de informação para o reconhecimento de gestos, enquanto que aqueles que pretendem reconhecer gestos em tempo real não são adequados para ambientes multicâmeras.

Dessa maneira, este trabalho propõe um sistema de reconhecimento de gestos *online* que possa ser utilizado em um espaço interacional multicâmera. Para isso, é proposta um abordagem semelhante à de Neverova et al. (2015); Chai et al. (2016), porém utilizando apenas os esqueletos 3D fornecidos pela implementação da proposta apresentada em Queiroz et al. (2018).

3. PROPOSTA

O reconhecedor de gestos proposto neste trabalho está dividido em três partes principais, as quais serão explicadas em seguida, a saber: extração de esqueleto 3D de um usuário, segmentação temporal do movimento, classificação dos gestos presentes nas sequências segmentadas.

3.1 Extração de Esqueleto 3D

Considerando um espaço inteligente com N câmeras, a extração de esqueletos utilizada será a abordada em Queiroz et al. (2018), como mencionado anteriormente. Dessa forma, a cada instante de tempo as imagens capturas pelas N câmeras são passadas para o reconstrutor de esqueletos que fornece um esqueleto 3D para cada usuário presente na cena. Cada coordenada 3D das juntas está no sistema de coordenadas no qual as câmeras foram calibradas.

3.2 Segmentação de movimento

Em uma aplicação *online*, as imagens são fornecidas ao modelo uma a uma, seguindo a ordem de captura. Assim, para que o reconhecedor possa classificar uma sequência de imagens como pertencente a uma das classes de gestos, é necessário um mecanismos de segmentação temporal que determine quando um possível gesto inicia e termina. O termo para esse tipo de modelo, como descrito em (Neverova et al., 2015; Chai et al., 2016), é *spotting*. Nesse sentido, propõe-se uma arquitetura dividida em dois principais componentes: um incorporador de características e uma rede neural recorrente multicamada.

Trabalhos como o de Neverova et al. (2015) extraem características específicas dos esqueletos, como angulações, velocidades e acelerações, a fim de melhorar a representação do movimento das juntas. Dessa forma, com o mesmo objetivo, propõe-se utilizar uma camada de incorporação de características (conhecida na literatura como *embedding*) baseada em uma rede neural com duas camadas totalmente conectadas, ambas com 32 neurônios de saída que ativam um função ReLU. Os pesos do *embedding* serão aprendidos durante processo de treinamento do segmentador temporal.

Para poder classificar um esqueleto como pertencente a um gesto ou a um não-gesto, é necessário levar em consideração a dependência temporal dos esqueletos classificados nos instantes de tempo anteriores. Dessa forma, propõe-se utilizar um classificador formado por duas LSTM empilhadas com 128 neurônios cada, seguidas por uma camada totalmente conectada com 2 neurônios de saída que ativam uma função *softmax*. A Figura 1 ilustra graficamente a arquitetura proposta para o modelo de *spotting*.

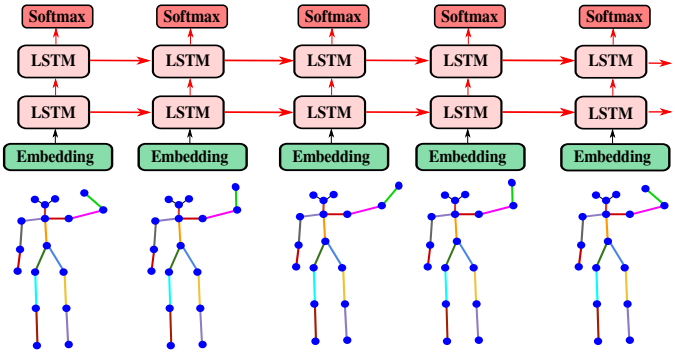

Figura 1. Arquitetura dos modelos usados na segmentação temporal do movimento e na classificação dos gestos.

3.3 Classificador

O classificador deverá receber uma sequência de esqueletos e classificá-la como pertencente a uma das possíveis classes de gestos. Dessa forma, a proposta é utilizar a mesma arquitetura do modelo de segmentação temporal (Figura 1), porém, modificando a última camada, de maneira que ela tenha tantos neurônios de saída, quanto o número de gestos a serem classificados.

4. EXPERIMENTOS

Nesta seção, será apresentado os experimentos que foram executados, o conjunto de dados utilizado, a configuração de cada experimento, como os resultados serão analisados e os recursos de *software* e *hardware* utilizados na implementação dos modelos.

O espaço inteligente utilizado neste trabalho está descrito em Almonfrey et al. (2018). Nele, existem quatro câmeras calibradas que se comunicam com o espaço através de uma infraestrutura física em rede e uma infraestrutura de *software* baseada em microsserviços. Para fins dos experimentos, considera-se que todos os usuários estão num mesmo plano do espaço.

4.1 Conjunto de dados

O conjunto de dados utilizado neste trabalho é composto por 1.200 vídeos (300 para cada câmera) que totalizam 2.400 gestos distribuídos entre 15 classes distintas. Cada vídeo corresponde a um gesto realizado oito vezes por

um mesmo voluntário. Como o espaço utilizado possui 4 câmeras, os voluntários foram instruídos a realizar cada um dos 15 gestos 8 vezes: de frente para cada câmera e de frente para um ponto entre elas. No total, foram realizadas 20 aquisições de 20 voluntários distintos em dois espaços inteligentes diferentes. No primeiro espaço inteligente, localizado no Instituto Federal do Espírito Santo, campus Vitória (Ifes-Vitória), foram realizadas 18 aquisições (2.160 gestos), as quais serão utilizadas para o treinamento dos modelos propostos. No segundo espaço, localizado na Universidade Federal do Espírito Santo, campus Goiabeiras, (Ufes - Lab Viros), foram realizadas mais duas aquisições (240 gestos), as quais serão utilizadas como conjunto de teste. Usar um conjunto de testes capturados em um outro espaço ajudará na validação da capacidade de generalização da proposta. Deve-se salientar que antes de cada aquisição de dados os voluntários assistiram a um vídeo demonstrativo de como cada gesto deveria ser executado. No entanto, mesmo executando um pouco diferente do indicado, as aquisições não foram interrompidas.

Todos os vídeos capturados foram rotulados de maneira que cada *frame* contenha o rótulo do gesto ao qual ele pertence: não-gesto (classe 0) ou gesto (classes de 1 a 15). Uma lista do vocabulário de gestos presentes no conjunto de dados é apresentada na Tabela 1.

Tabela 1. Lista dos gestos presentes no conjunto de dados utilizado. A classe 0 corresponde a um não-gesto, ou seja, quando nenhum gesto está sendo executado.

Classe	Gesto	Classe	Gesto
0	Não-gesto	8	Não (permissão)
1	Pedido de ajuda	9	Ruim (*feedback*)
2	Venha aqui	10	Dar passagem
3	Pode sair	11	Apontar
4	Siga-me	12	Dúvida
5	Pare	13	Mais alto (volume)
6	Abortar (missão/tarefa)	14	Mais baixo (volume)
7	Bom (*feedback*)	15	Silêncio

4.2 Pré-processamento dos dados

As imagens capturadas pelas quatro câmeras num mesmo instante de tempo passaram pelo processo de extração de esqueleto apresentado anteriormente. Sendo assim, foram gerados 300 sequências de esqueletos 3D (1.200 vídeos) correspondentes aos 2.400 gestos. Considerando os ruídos que podem ser gerados na etapa de extração, foi aplicada uma rotina de supressão de não máximos a fim de eliminar os esqueletos com juntas menos prováveis. Na Figura 2, pode ser vista a reconstrução de um esqueleto tridimensional a partir das imagens e dos esqueletos bidimencionais nelas detectados.

Como o modelo de esqueleto do *openpose* não possui a junta que representa o centro do torso (tratada aqui apenas como junta do torso), esta foi criada como sendo o ponto central do triângulo formado pelas juntas do tórax e dos quadris. Dessa forma, cada esqueleto foi centralizado na junta do torso e tiveram os comprimentos entre juntas normalizados de maneira a terem norma unitária. Em seguida, como as juntas dos membros inferiores comumente não trazem informações relevantes para a diferenciação dos gestos utilizados, optou-se por selecionar apenas sete

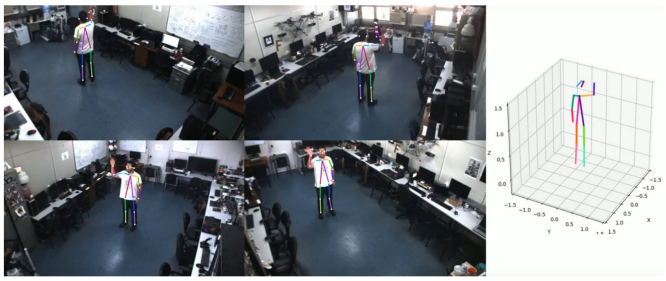

Figura 2. Exemplo da reconstrução de um esqueleto 3D a partir de quatro imagens adquiridas no espaço inteligente da Ufes.

juntas correspondente aos punhos, cotovelos, ombros e cabeça. Por fim, após a vetorização dos pontos das juntas, o vetor resultante foi normalizado para uma distribuição *Z-score* (média zero e variância unitária). Dessa forma, a entrada para ambos os modelos propostos é um vetor normalizado $\mathbf{e} \in \mathbb{R}^{1\times 21}$.

4.3 Treinamento

Os dois modelos propostos, o de *spotting* e o de classificação de gestos, foram treinados separadamente. Para isso, foi necessário utilizar todo o conjunto de treino, de forma que: (*i*) para o treinamento do *spotting* (classificador binário), as observações com rótulo 0 são consideradas não-gestos e as com rótulo maior que 0 são assumidas como gestos; (*ii*) para o treinamento do classificador, são utilizadas apenas as sequências segmentadas (aquelas que contêm gestos) e os correspondentes rótulos de classe.

A fim de escolher os hiperparâmetros a serem utilizados no treinamento, realizou-se uma validação cruzada de *10-folds* para cada modelo. Nesse caso, o conjunto de treinamento utilizado em cada um deles foi dividido em 10 partes, sendo nove utilizadas para o treinamento e um para a validação. Dessa forma, utilizou-se um processo de Otimização Bayesiana para determinar a configuração de hiperparâmetros que maximizasse a acurácia média das 10 rodadas de validação.

Após a obtenção do melhor conjunto de hiperparâmetros, treinaram-se os modelos utilizando o conjunto de treino completo. A Tabela 2 resume os hiperparâmetros utilizados no treinamento dos mesmos.

Tabela 2. Hiperparâmetros utilizados no treinamento dos modelos propostos.

Hiperparâmetros	Modelos	
	Spotting	Classificação
Tamanho da Sequência	256	32
Tamanho da subsequência	128	16
Número de épocas	100	300
Tamanho do lote	128	96
Taxa de aprendizagem	1e-3	1e-2
Taxa de decaimento L2	1e-5	1e-5
Otimizador	Adam	adam

4.4 Predição

Para a predição, utilizaram-se as sequências completas do conjunto de teste, possuindo gestos e não-gestos. Desse

modo, sempre que o modelo de *spotting* marcava uma observação como sendo gesto, as próximas observações eram armazenadas até que a primeira predição de um não-gesto ocorresse. Nesse momento, a sequência armazenada era passada para o modelo de classificação que determinava a qual das 15 classes de gestos a sequência pertencia.

É importante salientar que o modelo de *spotting* pode gerar vários ruídos na predição, o que prejudica a segmentação da sequência. Neste sentido, aplicou-se sobre a sua predição um filtro de média móvel com janela de tamanho 3 e sobreposição 2. Com isso, o problema pôde ser atenuado, uma vez que agora o *spotting* muda a sua predição de gesto para não-gesto quando a média é igual a 0, ou de não-gesto para gesto quando a média é igual a 1.

4.5 Método de avaliação

Os dois modelos serão avaliados conjunta e individualmente. Para o modelo segmentador, será utilizada a acurácia média calculada utilizando a classificação de cada observação. Já para o modelo classificador, será considerada a acurácia média da classificação de cada sequência contendo um gesto. Para o modelo completo, deve-se considerar que o segmentador pode detectar mais gestos que os existentes, ou simplesmente não detectar nenhum, o que acarretaria em problemas no uso da acurácia como métrica de avaliação.

A Figura 3 ilustra um exemplo onde a acurácia de classificação pode falhar. Nela o segmentador dividiu o gesto G_3 em dois gestos distintos (P_3 e P_4). Dessa forma, a sequência possui quatro gestos e o *spotting* identificou cinco. No caso de o classificador errar um dos outros gestos (G_1, G_2 ou G_4) e acertar os dois gestos preditos em G_3, a acurácia da classificação seria $1,0$, mesmo que um dos gestos da sequência não tenha sido classificado corretamente. Algo ainda pior seria se o classificador acertasse todas as predições (P_1, P_2, P_3, P_4 e P_5), o que daria uma acurácia de $1,25$, situação inadmissível, uma vez que a acurácia máxima que um modelo de classificação pode alcançar é de $1,0$.

Portanto, para avaliar o sistema completo, será utilizado o Índice de Jaccard (Equação (1)) que mede a sobreposição entre duas sequências, uma desejada (G) e outra predita (P), mediante o cálculo da interseção sobre a união entre as duas. Para isso, as observações classificadas como não-gestos recebem rótulo 0, aquelas classificadas como gesto e simultaneamente classificadas na classe correta recebem rótulo 0, caso contrário, as observações recebem o rótulo -1. O motivo de ter o rótulo -1 é fazer com que o erro de classificação não conte como um não-gesto. Nessa métrica de avaliação, o valor máximo de $1,0$ será alcançado apenas se ambos os modelos, segmentador e classificador, acertarem todas as predições.

$$J(G,P) = \frac{|G \cap P|}{|G \cup P|} \tag{1}$$

4.6 Detalhes de implementação

Todos os experimentos foram implementados em linguagem Python v3.7. Os modelos foram implementados utilizando o *framework* Pytorch v1.4. O processo de Otimização Bayesiana foi executado utilizando a biblioteca Hyperopt v0.2. O *Openpose* implementado em Tensorflow V1.2 foi utilizado na extração dos esqueletos das imagens. E, para a análise dos resultados, foram utilizadas as bibliotecas Matplotlib e Numpy.

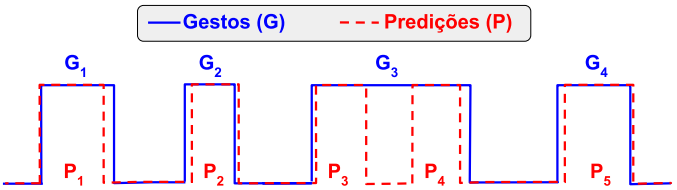

Figura 3. Exemplo do problema de avaliação. A sequência possui quatro gestos, mas foram detectados cinco.

O computador usado nos experimentos tem a seguinte configuração:

- sistema operacional Linux Ubuntu Server, distribuição 18.04;
- 2 processadores Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz com 12 núcleos físicos cada;
- 62 GB de RAM;
- 320GB de unidade de armazenamento (SSD);
- 4 placas gráficas (3 Nvidia Titan V e 1 Titan XP), com 12GB de memória cada.

Apesar do computador possuir quatro placas gráficas e dois processadores, os experimentos foram executados em apenas uma Titan V e em um dos processadores.

5. RESULTADOS E DISCUSSÕES

Após a realização dos experimentos, os resultados foram analisados e serão descrito na sequência.

O segmentador alcançou uma acurácia média de 82% sobre a base de testes. Considerando a média móvel utilizada, o segmentador pode estar realizando a predição três *frames* após o início e o fim do gesto, o que gera um erro de pelo menos seis observações por gesto. Dessa maneira, em um cálculo grosseiro, como o conjunto de teste possui 240 gestos e 18.932 observações (somadas as dos gestos e as dos não-gestos), tem-se um erro esperado de pelo menos 7%. Fazendo com que os 82% de acurácia alcançado seja ainda mais significativo. Isso mostra a efetividade do modelo, mesmo em um ambiente distinto e com usuários nunca vistos durante o treinamento.

Na Figura 4, é possível ver a matriz de confusão com os resultados do classificador. A acurácia média obtida na classificação dos gestos foi de $72,40\%$. Note que os gestos 1, 6, 10, 12 e 15 alcançaram a acurácia máxima de 100%. Os gestos 2 e 3 alcançaram quase 90%, enquanto que o pior resultado foi para o gesto 7, que alcançou apenas 6%. Mesmo com o baixo valor de acurácia para algumas classes, pode-se considerar que a proposta alcançou resultados satisfatórios. O vocabulário de gestos utilizado possui gestos ambíguos que dependem não só do movimento dos braços, mas também da forma da mão. Dessa maneira, os gestos que podem ser diferenciados apenas com o movimento alcançaram a acurácia máxima. Por outro lado, aqueles mais dependentes da forma da mão foram confundidos uns com os outros, como já era de se esperar.

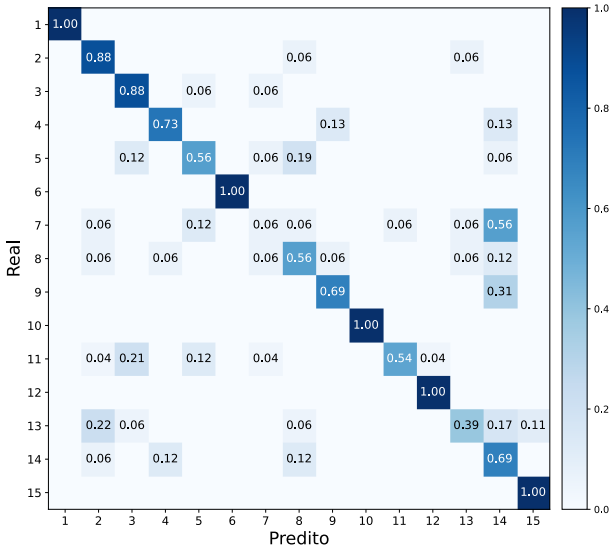

Figura 4. Matriz de confusão com os resultados do classificador de gestos.

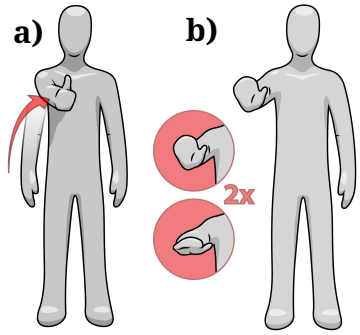

Figura 5. Exemplo de dois gestos que podem ser confundidos. a) gesto 7 e b) gesto 14. Nessas duas classes de gestos, o que as difere é a forma da mão.

Como exemplo, o que diferencia o gesto 7 do gesto 14 (Figura 5) é a forma das mãos, uma vez que eles possuem o mesmo movimento. No gesto 7, a mão está fechada e com o polegar apontando para cima, enquanto que no gesto 14 a mão está aberta com palma inicialmente para cima e depois para baixo.

Considerando a operação conjunta de ambos os modelos, o índice de Jaccard obtido foi de $0,76$. Ao tomar em conta todas as restrições intrínsecas ao vocabulário de gestos utilizado, julga-se que esse valor também é satisfatório. Implicando que o modelo de *spotting* e classificação possa ser utilizado em um ambiente interacional *online*. A Figura 6 traz o exemplo de um trecho com as predições e o rótulo utilizado na Equação (1).

Um outro resultado importante é trazido na Tabela 3. Nela, podem ser vistos os tempos de execução demandados pela solução quando executada em GPU e CPU. Utilizando GPU, cada observação pode ser processada em 70ms, em média. Já em CPU, ela chega a ultrapassar os 1.200ms. Com isso, utilizando GPU é possível ter um reconhecedor de gestos que responde a uma aplicação em menos de 100ms, ou seja consegue operar a pelo menos a 10 FPS. Importante notar que o modelos reconhecedor (*spotting* e classificador) demanda praticamente o mesmo

Tabela 3. Tempos médios de execução para cada observação.

	Tempo médio (ms)		
Dispositivo	*Esqueleto*	Reconhecedor	Total
CPU	1.262	0,6	1.263
GPU	69	0,4	70

tempo de processamento, seja em CPU, seja em GPU. Dessa forma, ele pode ser executado em CPU sem que haja perdas significativas no tempo de resposta da aplicação.

5.1 Comentários gerais

Apesar dos resultados alcançados pelos modelos propostos serem promissores, notam-se algumas limitações que necessitam ser abordadas. A informação da forma das mãos não pôde ser extraída do esqueleto fornecido, e esse tipo de problema é enfrentado por todos os modelos que utilizam o esqueleto como fonte de informação, porém, raramente isso é discutido. Dessa forma, resultados mais significativos só poderão ser alcançados caso seja possível levar em consideração a forma das mãos. Apesar do *openpose* possibilitar extrair as juntas dos dedos das mãos, ele necessita que a mão esteja nítida, o que não acontece nesse tipo de ambiente. A distância do usuário para cada câmera é de pelo menos 3 metros. Dessa forma, é necessário utilizar outro tipo de abordagem para a extração das características das mãos. Outro aspecto importante é que, para as quatro câmeras, têm-se ao menos oito mãos a serem consideradas a cada instante de tempo. Decidir qual câmera utilizar também é de primordial importância, principalmente quando o número de câmeras aumenta significativamente.

Num espaço inteligente, é comum haver múltiplos usuários interagindo ao mesmo tempo. Dessa maneira, cada um pode realizar gestos diferentes a cada instante de tempo. Nesse cenário, qual a abordagem deveria ser usada: um modelo para cada usuário, ou apenas um modelo que considera todos os usuários? Para ambas as respostas, é necessário primeiro que esqueleto seja identificado unicamente em todas as observações, o que é um outro problema ainda em aberto na literatura. Soluções para essas duas perguntas devem ser cuidadosamente analisadas, para que o sistema não seja sobrecarregado pela quantidade de modelos em execução simultaneamente ou pelo processamento demandado por um único modelo que tenta resolver todo o problema.

6. CONCLUSÕES E TRABALHOS FUTUROS

Os espaços inteligentes multicâmeras possibilitam uma maior cobertura do ambiente e a redução de problemas de oclusão. Para interagir com tais espaços, é desejável que a interface de interação seja intuitiva, sendo os gestos dinâmicos uma das mais indicadas.

O problema de utilizar gestos dinâmicos em ambientes multicâmeras é o aumento da complexidade dos reconhecedores. Além do mais, nesse tipo de ambiente, é imprescindível que o reconhecedor funcione de modo *online*. O que aumenta ainda mais os seus requisitos.

Com vista a contribuir na solução para esse problema, este trabalho propôs um reconhecedor de gestos *online*

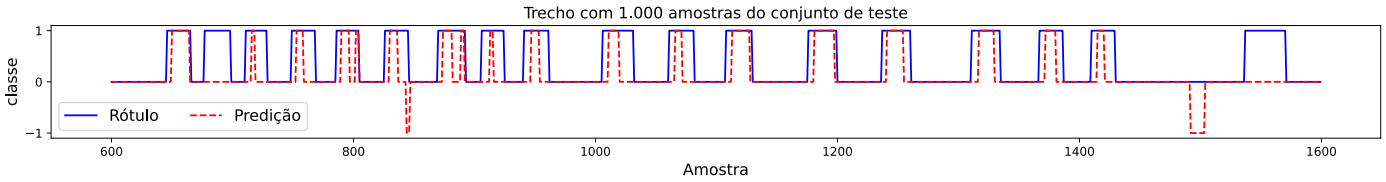

Figura 6. Exemplo de um trecho da base de teste utilizado no cálculo de índice de Jaccard (Equação (1)). Na predição, a classe 1 significa que o gesto foi classificado corretamente, a classe 0 é um não-gesto e a classe −1 corresponde a um gesto detectado, porém erroneamente classificado. Para o rótulo, a classe 1 corresponde a um gesto e 0 a um não-gesto.

para um ambiente multicâmera baseado em esqueleto 3D. Para isso, foi utilizado um método para reconstruir o esqueleto a partir de imagens de câmeras calibradas, um modelo capaz de segmentar temporalmente as sequências contendo gestos e um outro modelo que classifica os gestos correspondentes a tais sequências.

Como resultado, o modelo segmentador obteve uma acurácia média sobre o conjunto de dados de teste de 82% e o modelo classificador obteve 72, 40%. Os dois modelos em conjunto alcançaram um índice de Jaccard de 0.76. Considerando o tempo de processamento, cada observação requereu um tempo médio de 70ms quando executado em GPU. Dessa forma, julga-se a solução proposta efetiva no reconhecimento *online* de gestos dinâmicos para ambientes interacionais multicâmeras.

Apesar de satisfatórios, alguns trabalhos ainda precisam ser feitos a fim de alcançar resultadas mais significativos. Primeiro, é necessário que o modelo inclua a informação da forma das mãos para que os gestos com movimentos ambíguos possam ser reconhecidos. Segundo, é necessário que o sistema seja capaz de reconhecer gestos de múltiplos usuários. Para isso, deve-se determinar como os modelos devem tratar as informações dos diversos usuários e como estes serão relacionados a cada nova observação.

REFERÊNCIAS

Almonfrey, D., do Carmo, A.P., de Queiroz, F.M., Picoreti, R., Vassallo, R.F., and Salles, E.O.T. (2018). A flexible human detection service suitable for Intelligent Spaces based on a multi-camera network. *International Journal of Distributed Sensor Networks*, 14(3), 155014771876355. doi:10.1177/1550147718763550.

Cao, C., Zhang, Y., and Lu, H. (2015). Multi-modal learning for gesture recognition. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., and Sheikh, Y. (2018). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*.

Chai, X., Liu, Z., Yin, F., Liu, Z., and Chen, X. (2016). Two streams recurrent neural networks for large-scale continuous gesture recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 31–36. IEEE.

Escobedo-Cardenas, E. and Camara-Chavez, G. (2015). A robust gesture recognition using hand local data and skeleton trajectory. In *2015 IEEE International Conference on Image Processing (ICIP)*, 1240–1244. IEEE.

Girshick, R. (2015). Fast r-cnn. In *International Conference on Computer Vision (ICCV)*.

Lee, J.H., Ando, N., and Hashimoto, H. (1999). Intelligent space for human and mobile robot. In *Advanced Intelligent Mechatronics, 1999. Proceedings. 1999 IEEE/ASME International Conference on*, 784–784. IEEE.

Liu, X., Shi, H., Hong, X., Chen, H., Tao, D., and Zhao, G. (2019). Hidden states exploration for 3d skeleton-based gesture recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1846–1855. IEEE.

Neverova, N., Wolf, C., Taylor, G., and Nebout, F. (2015). Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8), 1692–1706.

Queiroz, F.M., Picoreti, R., Santos, C.C., Fernandes, M.R., and Vassallo, R.F. (2018). Estimating tridimensional coordinates of skeleton joints in a multicamera system. In *Anais do XIV Workshop de Visão Computacional - WVC 2018*, 108–114.

Saleme, E.B., Celestrini, J.R., and Santos, C.A.S. (2017). Time evaluation for the integration of a gestural interactive application with a distributed mulsemedia platform. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, 308–314.

Santos, C.A.S., Neto, A.N.R., and Saleme, E.B. (2015). An event driven approach for integrating multi-sensory effects to interactive environments. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 981–986. IEEE.

Santos, C.C., Picoreti, R., do Carmo, A.P., Vassallo, R.F., and Garcia, A.S. (2017). Modelagem de um comportamento interacional entre homem e robô para um espaço inteligente baseado em visao computacional. *Simpósio Brasileiro de Automaçao Inteligente, SBAI*.

Santos, C.C., Samatelo, J.L.A., and Vassallo, R.F. (2020). Dynamic gesture recognition by using cnns and star rgb: A temporal information condensation. *Neurocomputing*.

Santos, C.C.d. et al. (2016). Proposta de uma metodologia para a obtenção de vocabulários de gestos intuitivos para a interação homem-robô.

Wang, H. and Wang, L. (2017). Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 499–508.

Zhao, C., Pan, W., and Hu, H. (2013). Interactive indoor environment mapping through human-robot interaction. *International Journal of Modelling Identification & Control*.